# Early Y chromosome lineages in Africa: the origin and dispersal of *Homo sapiens*

Chiara Batini (1,2)†, Gianmarco Ferri (3), Giovanni Destro-Bisol(2, 4), Francesca Brisighelli (4,5),Donata Luiselli (6), Paula Sanchez-Diz (7), Jorge Rocha (8), Lynn Jorde (9), Antonio Brehm (10), Valeria Montano (2), Nasr Eldin Elwali (11) Gabriella Spedini (2,4), Maria E. D'amato (12), Natalie Myres (13), Peter Ebbesen (14), David Comas (1), Cristian Capelli (5)*

*(1) Institute of Evolutionary Biology (UPF-CSIC), CEXS-UPF-PRBB, Barcelona, Spain*

*(2) Dipartimento di Biologia Animale e dell'Uomo, Sapienza Università di Roma, Italy*

*(3) Department of Diagnostic and Laboratory Service and Legal Medicine, Section of Legal Medicine, University of Modena and Reggio Emilia, Italy*

*(4) Istituto Italiano di Antropologia, Roma, Italy*

*(5) Department of Zoology, University of Oxford, UK*

*(6) Dipartimento di Biologia Evoluzionistica Sperimentale, Unità di Antropologia, Università di Bologna, Bologna, Italy*

*(7) Institute of Legal Medicine, Genomics Medicine Group, University of Santiago de Compostela, Spain*

*(8) Instituto de Patologia e Imunologia Molecular (IPATIMUP), Universidade do Porto, Portugal*

*(9) Department of Human Genetics, University of Utah Health Sciences Center, Salt Lake City, UT 84112*

*(10) Human Genetics Laboratory, University of Madeira, Campus of Penteada, Funchal , Portugal*

*(11) Department of Molecular Biology, National Cancer Institute (NCI-UG), University of Gezira, Wad Medani, Sudan.*

*(12) University of the Western Cape, Department of Biotechnology, Forensic DNA Lab, Cape Town, South Africa.*

*(13) Sorenson Molecular Genealogy Foundation, Salt Lake City, Utah, USA*

*(14) Department of Health Science and Technology, Aalborg University, Denmark*

† *Current address: Department of Genetics, University of Leicester, UK*

**Corresponding author (cristian.capelli@zoo.ox.ac.uk)*

**The study of Y chromosome variation in extant populations has provided significant insights into the genetic history of *Homo sapiens* (Jobling and Tyler-Smith, 2003; Soares et al, 2010; Stoneking & Delfin, 2010; O'Rourke & Raff, 2010). Focusing on sub-Saharan Africa, demographic events associated with the spread of languages, agriculture and pastoralism have been targeted (Destro-Bisol et al, 2004; Beleza et al., 2005; Wood et al, 2005; Tishkoff et al, 2007; Berniell-Lee et al., 2009; Henn et al, 2009; Cruciani et al., 2010), although little is known about ancient population history. The first two branches of the Y chromosome genealogy, namely haplogroup A and B, are African specific, with average continental frequencies of 14-34%, reaching up to 65% in groups of foragers (Cruciani et al, 2002; Wood et al, 2005; Tishkoff et al, 2007; Berniell-Lee et al, 2009). Despite the potential of such lineages in revealing signatures of the ancient peopling of the continent, an exhaustive investigation of their distribution and variation is currently missing. Here we show that their systematic dissection provides novel insights into the early history of our species. We highlighted both a complex pattern of populations' dynamics associated with the Last Glacial Maximum, shaping the separation among hunter-gatherer communities and the peopling of Western and Southern Africa, and the retention of the very early human Y chromosome lineages in Eastern and Central but not Southern Africa. These results open new perspectives on the early African history of *Homo sapiens*, with particular attention to areas of the continent where human fossil remains and archaeological data are scanty.**

Sub-Saharan African Y-chromosome diversity is represented by five main haplogroups (hgs): A, B, E, J and R (Underhill et al, 2000; Cruciani et al, 2002; Tishkoff et al, 2007). The most common hg is E, whose distribution across the continent parallel that of Bantu-speaking communities, while hgs J and R are geographically restricted to East and Central Africa, respectively (Underhill et al, 2000; Cruciani et al, 2002; Tishkoff et al, 2007; Berniell-Lee et al, 2009; Cruciani et al, 2009). The other two groups, A and B, are dispersed across different geographic areas and populations, suggesting an association with complex, potentially more ancient, demographic events (Underhill et al, 2000; Cruciani et al, 2002; Tishkoff et al, 2007).

We screened literature data and genotyped both novel and previously partially investigated samples for a total of approximately 10,000 chromosomes from more than 180 populations (Table S1), collecting data on 185 hg A and 457 hg B Y chromosomes (Table S2). Outside Africa, such chromosomes have been sporadically found in Europe and America (Semino et al, 2000; Luis et al, 2004; Capelli et al, 2006; Hammer et al, 2006; King et al, 2007; Jim Wilson, personal communication). Hg A is rarely found in North, Western and Central Africa while is more frequent in the Eastern and Southern parts of the continent. Absent from the North and rare in Western Africa, Hg B distribution in the rest of the continent mirrors that of its two main sub-clades B2a and B2b (Figure 1). The former appeared to be associated with food-producing communities and populations in contacts with those (as shown also for hg E; Beleza et al., 2005; Berniell-Lee et al., 2009) while B2b is mostly present in foraging communities in Eastern and Central Africa. Hg A and B different geographic distributions are mirrored at population level (table S3). Little or no hg A is present in Pygmies and Eastern Africa (EA) Khoesan speakers (for the use of the word Khoesan and issues with populations classification in Southern Africa see Mitchell, 2010), while B2b is commonly found in these populations. On the other hand, hg A is more frequent than B among Southern Africa (SA) Khoesan speakers, with B2a and B2b representing approximately 18% of the Y chromosome types present in these populations (Table S2).

Diversity indexes are shown in table 1. Overall, hg A shows higher diversity than B and, within the latter, B2a presents lower values than B2b.The evolutionary relationships among haplotypes within these hgs are shown in Figure 2 and figure S1. A clear correlation between genealogy and geography/populations

was evident for hg A and B2b sub-lineages (Figure 2a, b) while hg B2a network analysis revealed a star-like structure, as expected in the case of relatively recent dispersion events (Figure S1; Beleza et al., 2005; Berniell-Lee et al., 2009). Hg A1 is found only in Western and Central Africa, while A3b2 and A3b1 are Central/Eastern and Southern African specific respectively. Hg A2 is mostly represented by Southern African (SA) samples, with only few central African haplotypes. Similarly, B2b1/B2b4a and B2b2 are geographically restricted to SA and East Africa, respectively. Such structured geographic distribution is also present at population level. B2b3, B2b4b and B2b4* (as well as the previously un-described MSY2* lineage; Figure S2b) are all almost exclusively found among Western Pygmies, while B2b2 and B2b1-B2b4a are found only in Eastern Pygmies and SA Khoesan speakers, respectively. Similarly, the majority of the A3b1 and A2 types are found among SA Khoesan speakers, with hg A2 also present in Western Pygmies. Pygmies and SA Khoesan speakers have evolutionary close lineages also within B2b4 clade (Wood et al, 2005; this work).

Our extensive survey of SNP variation in A and B Y chromosomes enabled us to detect several genealogical incompatibilities with the recently proposed topology (Figure S2, Karafet et al, 2008). PK1, originally thought to be an hg A2 specific marker, resulted shared between A2 and A3 and new lineages have been identified on the long branch characterising A2 (Fig. S12). M190 had been indicated as specific to A3b clade, but in our analysis it resulted derived in all A3 lineages. P7 appears to be basal to most of B2b lineages. Within the P7 derived chromosomes, the MSY2 marker clusters lineages defined by M211, M115/M169, M30/M129 and P8/P70 (Fig. S2b). The identification of two chromosomes derived at the above markers but not for P7 suggests that this polymorphism might be prone to recurrent mutations, possibly due to gene conversion (Mark Jobling personal communication) (Figure S2b). For simplicity, we have retained the same nomenclature as recently described in Karafet et al, 2009 (with the exception of MSY2*, see above) but in future renaming would be necessary.

In order to provide a temporal frame to the observed pattern of distribution of genetic diversity, we estimated microsatellite based Average Squared Distance (ASD; Goldstein et al, 1995) between and within clades (see Methods section). We specifically focused on lineages providing insights on the peopling of Western and Southern Africa (A1, A2 and A3b1), the relationship between Western and Eastern Pygmies (B2b3, B2b4b and B2b3) and between Khoesan and Pygmies (B2b4 and A2) (table 2).

Results are consistent with an early split within A1 clade approximately 32.7 thousand years ago (Kya) (95% confidence intervals - CI - 24.8-38.1 Kya; Table 2a) with a western Africa specific diversity dating back to 16.6 Kya (CI 6.6-38.7 Kya; table 2b), in agreement with early archaeological and linguistic evidence. The Ounanian culture has been recorded in Mali as far back as 9-10Kya (Clark, 1980; Raimbault, 1990; MacDonald, 1998) and the lithic and ceramic assemblages from Ounjougou dates back to 12Kya (Huysecom et al, 2004; 2009). Similarly, the origin of the early Niger-Congo Atlantic branch has been placed at least 8 Kya (Ehret, 2000; Blench, 2006). The detection of a genetic signal associated with ancient human presence in this area is of interest given the homogeneity between Western and Central African populations that has been observed so far for other Y chromosome lineages (hg E for example) and genome wide analysis (Li et al, 2008; Tishkoff et al, 2009).

The hg A2 SA specific variation dates back to 14.2 Kya (CI 5.3-35Kya), while A3b1 estimated date (the other SA specific clade) is 23.5Kya (CI 10.2-52 Kya). The two sister clades A3b1 and A3b2 have overlapping temporal ranges, with a separation time of 39 Kya (CI 31.6-43.4 Kya). A3b2 is mostly frequent among Nilo-Saharan populations from Sudan and Kenya, suggesting an association between the two, as recently proposed by Gomes et al, 2010. The initial date for the spread of this linguistic phylum, approximately 18,000 years ago (Blench, 2006), is in accordance with the estimated within-clade A3b2 date (18.9 Kya, CI 7.6-41.9 Kya).

Eastern-Western Pygmies separation has been dated using the divergence between B2b2 and B2b4b/B2b3 clades (table 2a). These estimates do not overlap (28.7 Kya, CI 24-32.8 Kya; 42.3 Kya, CI 34.7-54.5 Kya) but the smaller of the two provides an indication of the lower bound of such separation (B2b2 vs B2b4b). The time of this split is consistent with results based on mtDNA and autosomal loci (Destro-Bisol et al, 2004; Patin et al, 2009) and in line with the habitat fragmentation of tropical forest that followed the Last Glacial Maximum. We also noted that the within clade diversity/antiquity is extremely reduced for these Pygmy-specific lineages, suggesting a bottleneck in the relatively recent demographic history of these groups, as observed for other loci (Excoffier e Schneider 1999; Patin et al, 2009; Batini et al, submitted).

Unexpectedly, we noted the presence of B2b4* chromosomes in Mozambican samples, a lineage only shared with Western Pygmies. The low frequency of these chromosomes in the SE African populations,

together with the lack of appropriate evidence of a link among early inhabitants of south-east Africa with Western Pygmies leave the issue difficult to disentangle and call for a more detailed and focused investigation. In this sense, a worth exploring scenario could be based on the presence of this lineage in pre-Bantu populations already settled in the regions, which could have been absorbed by the incoming agro-pastoralists groups (Sikora et al., submitted).

A3b1 and A2 dates extend beyond the Last Glacial Maximum (LGM; 25-18Kya) in line with human Late Stone Age presence in Southern Africa (Mitchell, 1996) but not as early as suggested by fossil and archaeological remains (White et al, 2003; McDougall, Brown and Fleagle, 2005; Lewin and Foley, 2005; Marean et al, 2007). Our somehow more recent estimates for SA hgs A2 and A3b1 (table 2b) could be possibly due to our partial population coverage, which might be tested through more extensive population surveys (Quintana-Murci et al, 2010; Marks S and Capelli C, unpublished results on Lesotho populations) as well as past lineage extinction that followed the significant demographic changes that occurred during the Marine Isotope Stage 3 (60-25 kya) and the Last Glacial Maximum (Mitchell et al, 2008). It is also worth considering the possibility that A2 and A3b1 retain signatures of two independent pre-Bantu dispersal events. This scenario is somehow also supported by the different geographic distribution of these two clades. A3b1 is present all across southern Africa while A2 is found almost exclusively associated with populations in South-Western Africa or originally from this area (see table S3).The A2 distribution broadly overlap that of Khoe languages and it could potentially represent a genetic signature of the contacts/migrations of the khoekhoen pastoralist societies from Northern Botswana, Southern Angola and western Zambia area approximately 2kya (Mitchell and Whitelaw, 2005; see also below).

We pinpointed novel evolutionary links between Western Pygmies and SA Khoesan speakers, which develop previous findings of by Wood et al, (2005). Hg A2, commonly found among SA Khoesan speakers (6-45%), was detected at non trivial frequency (5%) among the Baka Pygmies. On the other hand, B2b4 is present at 6-7% among Khoesan speakers but reach 45-67% in both Biaka and Baka pygmies (Table S3). We dated to 10-12Kya the TMRCA among Western Pygmies and Khoesan groups for these two haplogroups, with the CI arguing against a role of the Bantu-speaker expansion (CI 4.4-16 Kya; 5.4-18.6 Kya; table 2c). Evidence for a Pygmy /Khoesan link has been provided by recent genome wide studies. In the re-analysis by Hellenthal et al, 2008 of data from Conrad et al, (2006) the first link to emerge among

populations is indeed between San and Western Pygmies. Furthermore, Tishkoff et al, 2009 have recently observed shared ancestry between San and Eastern Pygmies, and more broadly, also with Western Pygmies and Hadza from Tanzania. Intriguingly, the genetic link seems to be paralleled by a sort of cultural "synapomorphy" such as that implied by the connection between the Batwa (considered related to modern Pygmies, Smith 1995; 1998; 2008) and the Khoekhoen groups possibly on the basis of shared rock art geometric design (Smith and Ouzman, 2004).

Although B2a has not been investigated with the same resolution of A and B2b hgs, our data support an association with Bantu-speaking populations (Beleza et al, 2005; Berniell-Lee et al, 2009; see table S1). Within-clade variation suggests a more recent origin for B2a than B2b, while Network analysis did not reveal population/geography specific STR-based clusters (Figure S1). Further analysis within the B2a clade with a deeper phylogenetic resolution together with additional populations may help to clarify the demographic dynamics associated with its dispersal.

Whereas the dissection of single Y-chromosomal clades or sub-clades proved useful to shed light on the relations between specific populations/groups and helped reconstruct the demographic impact of migratory and cultural events, a wider and exhaustive phylogeographic analysis may provide indications on areas of the African continent where the extant human Y-chromosome diversity first originated.
The haplogroups A and B are ideal candidates for this task, given their distribution in Africa and the fact that they represent the earliest lineages to branch off within the Y chromosome genealogy. Previous analysis pointed to SA EA following the identification of hg A types (A3b1 and A3b2) in populations from these areas (Hammer et al, 2001). However, our results clearly indicate that A3b branched later within hg A, making it not informative on the origin of the early human Y lineages. Hg A is divided in two branches: A1, represented by West and Central Africa types, and A2-A3, containing South and Eastern Africa chromosomes, plus few from Central Africa. A2 is mostly composed by Southern Africa types; however, an early branch in A2 is from Central Africa. Within A3, A3b1, the southern Africa clade, is a sister clade of A3b2, common in Eastern Africa, while A3a is found among Eastern Africans only (figure XX). In B, B2a and B2b are two sister clades, while B* groups a number of chromosomes from Central Africa ancestral for the set of SNPs we tested. B2a has a very wide distribution, possibly associated with the relatively recent dispersion of Bantu-speaking populations from Central Africa. Within B2b, B2b* contains samples from

Eastern, South-Eastern and Central Africa, with P6 derived chromosomes from South Africa and P7 types are mainly from hunter-gatherer populations from Central, Eastern and Southern Africa (see Figure 2c). These results together with the genealogical depths of Southern Africa specific clades are suggestive of this area being an early recipient of ancient human migrations from other regions more than the original source, in contrast with recently published autosomal data (Tishkoff et al, 2009). In respect of Eastern and Central Africa roles, the dataset we presented, while tentatively pointing to a wider preservation of ancient lineages in Central Africa, is still compatible with a primary role for Eastern Africa, a hypothesis suggested also by mtDNA analysis (Behar et al, 2009). We note that the current absence of significant palaeo-anthropological investigation coupled with the different possibility of fossil preservation in Central Africa makes the extremely long human fossil record in Eastern Africa not conclusive in solving this issue. The screening of Y-chromosomal variation at the same level of resolution in additional populations from these regions as well as the analysis of genomic data is expected to provide further details on the early steps of *Homo sapiens* in Africa.

**Acknowledgments**

We would like to thank

Sergio Tofanelli and Davide Merlitti for giving access to early versions of the ASHEs software; Jim Wilson, Fabio Verginelli and Renato Mariani-Costantini for providing samples and unpublished data; Peter Mitchell for helpful discussions on the African archaeological record; Marco Giorgi for providing scripts used during data analysis. MORE Mònica Vallés, Stéphanie Plaza, and Roger Anglada (UPF) for technical support.

CC is a RCUK Academic Fellow. Spanish ministry fellowship

## METHODS

### Y chromosome SNP/STR genotyping

Samples have been genotyped with different set of markers (Table S2). SNP scoring has been conducted using mini-sequencing multiplex reactions and direct sequencing (Supplementary Material). STR genotyping was conducted using commercially available STR kits (Krenke et al, 2003; Mulero et al, 2006) as well as in-house developed multiplexes. Almost all the samples included were genotyped for 10

microsatellites: DYS19, DYS389-I, DYS389II- estimated by subtracting the DYS389 alleles-, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439. A subset of the samples was tested for additional 5 loci (DYS448, DYS456, DYS458, DYS635 and Y-GATA-H4). In the statistical analysis two sets were used: one based on 8 of the 10 STRs present in the majority of the samples (excluding DYS389II and DYS390 due to their complex repeat structure) and another including all the 15 STR markers. The bi-allelic locus DYS385 was not included in any of the analysis for uncertainty in locus assignation of the two alleles.

**Network reconstruction and diversity estimation.** Median-Joining networks of microsatellite haplotypes have been constructed using Network 4.5 and Network Publisher. Weights have been estimated using the inverse of the individual loci within clade variances for STRs. SNPs have been weighted according to their hierarchical position in the genealogy identified in the present paper. Within-hg diversity was estimated using several indexes. Total number of haplotypes, unique haplotypes and haplotype diversity were estimated using Arlequin (Excoffier et al, 1995). The variance was estimated as the within–locus mean allele variance averaged across all loci. C.I. were based on 10,000 resampling performed across individuals. The pattern observed using 15 or 8 STRs was similar (data not shown). Given the substantially larger (59% and 149% more chromosomes for A and B respectively) and geographically more comprehensive datasets available using the smaller set of STRs, we reported the estimates based on 8 microsatellites (Table 1).

**Dating.** Haplogroup genealogical depths have been estimated using the model-free statistics average squared difference (ASD; Goldstein et al, 1995) (Table 2b). The TMRCA (Time to the Most Recent Common Ancestor) of a clade was estimated by calculating ASD between all current chromosomes of that lineage and the founder haplotype reconstructed by combining the modal alleles at single loci (Thomas et al, 1998). The ASD estimated in this way has an expected value of $\omega T$, where $\omega$ is an average effective mutation rate at the loci, taken as $6.9*10^{-4}$ per 25 years (Zhivotovsky et al. 2004) and T is the separation time expressed in number of generations. This approach is expected to underestimate the age of the clade if the reconstructed founder haplotype differs from the real one. 95% confidence intervals were estimated using the program Ytime (Behar et al, 2003). The TMRCA estimated with this approach can be used as an indication of the lower bound of the split among different groups. Comparing different datasets on the basis of the considered number of loci, two parameters appears to influence the TMRCAs estimates: sample size and number of loci. The impact of the two parameters on the estimates is not easy to predict as the majority

of the investigated lineages are strongly structured and the number of STRs genotyped in not uniformly distributed within our dataset. It follows that the selection of a high number of STRs (15) sometimes lead to the exclusion of significant portion of the diversity present within haplogroups (see above).. Taking in consideration that when the within lineage branching coverage did not significantly change by using either 15 or 8 STRs haplotypes the estimates are substantially overlapping (data not shown), we decided to use the largest dataset available for our founder dating estimates, using haplotypes defined by 8 STRs (Table 2b). It should be also noted that the many of these lineages are particularly rare and that the within clade variation might have been only partially surveyed, a condition that might be skewing current estimates towards the lower bound of the real genealogical depth.

The upper bound of lineage split can be estimated using ASD calculated between lineages. (ASD=2$\mu$T; Goldstein et al, 1995). ASD is based on a strict single stepwise mutation model (Goldstein et al, 1995). However, in the presence of multi-step mutational events the squaring process is expected to heavily influence the distance estimation, corrupting the linearity with time. In order to take into account such occurrences and limit the impact of multi-step mutations, we compared the average number of mutational steps across loci. The average and the variance of these estimates were used to identify outliers: if the number of observed steps were outside the 95% CI, the associated locus was removed. We sequentially applied this approach until all loci were within the 95% CI. Using this approach we identified potential deviation from the single-stepwise model comparing B2b2 vs. B2b3 and B2b4b at locus DYS19, A1-M31 vs. A1-P108xM31 at loci DYS390, DYS392, DYS438 and DYS635 and at loci DYS635 and Y-GATA for A3b1 vs. A3b2. These loci were excluded from calculations and ASD was estimated averaging across the remaining 13 loci (11 for A1 comparison). All available STRs/haplotypes were initially considered (DYS19, DYS389-I and DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and Y-GATA-H4). Haplotypes with missing data and loci suggesting departure from a strict-stepwise mutation model (see above) were removed. 95% confidence intervals were estimated by calculating 10,000 ASD values per pair of haplogroups based on re-sampled individuals. To evaluate the performance of this approach, we also estimated the ASD between the A and B lineages. We note that their TMRCA can be equated to the depth of the Y chromosome genealogy. Using the same procedure as described above we included 15 STRs and estimated the average ASD comparing main lineages within A

and B( n>10; A1, A2, A3b1, A3b2, B2a, B2b) and obtained a TMRCA for A and B of 64.2Kya, in agreement with other estimates based on sequence data (70Kya; Thomson et al, 2000).

TMRCAs between a single haplotype and a group of chromosomes (as the case for the Khoesan speakers-Pygmies comparisons in clades A2 and B2b4) was calculated using the approach described by Walsh, 2001 as implemented in the software ASHEs (Tofanelli et al, 2009). Briefly, TMRCA Bayesian Posterior distributions are calculated for pair of chromosomes separated by n mutational steps, assuming a step-wise mutational model, an average mutation rate of $6.9*10^{-4}$ per locus per generation, a lambda value of 0.0002 (1/N, where N=population size; here used N= 5,000 in accordance to other investigations, Hammer et al, 1995; Walsh, 2001) investigated over 3,000 generations (given the depth of the Y chromosome genealogy of approximately 70Ky, assuming a generation time of 25 generations, such value can be considered as appropriate to fully explore the likelihood distribution). The most likely TMRCA estimation for each pair of chromosomes were identified and used to calculate the average TMRCA within each set of comparisons. To estimate the 95% boundaries, we approximated each distribution to a normal one, centred on the most likely TMRCA value. We then calculated the area below the posterior distribution comprised between the most likely value and the beginning of the curve, and equated this value to 50% of the total likelihood curve. The 2.5% lower bound of the distribution was identified as the TMRCA value comprising the 95% of the right hand side likelihood of the curve. By symmetry, the value for the upper bound was similarly estimated (table 4c). The 95% CI reported in table 4c are the averages across comparisons based on each 2.5 and 97.5% estimates. As previously shown (Walsh, 2001), the robustness of such estimates is strongly influenced by the number of STRs investigated, so we tried to maximise the number of investigated loci. However, fewer microsatellites (i.e. 7 for example) tended to provide younger estimates still overlapping with the dates obtained with more loci (data not shown). We tested for the presence of possible multistep mutations as described above and removed the related loci from the datasets (DYS390 and DYS389II when comparing B2b4 and A2 haplotypes respectively). The values shown in table 4c are in agreement with the ASD estimates based on the same datasets (central values: 10.2Kya and 9.2 Kya, respectively; data not shown)

References

**Legend to Figures and Tables**

Fig. 1 – Frequencies of A and B in Africa sub-regions. yellow represent hg A, green B2a and blue B2b.
Fig. 2 – Genealogical relations among A and B SNP-based and STR-based haplotypes. a) A haplogroup, geographical colour scheme; b) B2b haplogroup, geographical colour scheme; c) A and B SNP-based haplotypes, geographical colour scheme
Fig. S1 – Genealogical relations among B2a STR-based haplotypes.
Fig. S2 – Genealogical relations among SNP-based A (a) and B2b (b) haplotypes – comparison with Karafet 2008

Tab. 1 – Diversity indexes for hg A and B, including sub-haplogroups B2a and B2b, based on 8 STRs (DYS19,DYS389I DYS391, DYS392, DYS393, DYS437, DYS438, DYS439). N:total number of individuals; K:total number of haplotypes; Ks:total number of unique haplotypes. The 95% C.I. for the variance were estimated by 10,000 resampling across individuals
Tab. 2 – Dating (a, b and c)

Tab. S1 – database and frequencies
Tab. S2 – haplotype list (A, B2a and B2b)
Tab. S3 –A and B2b sub-lineage distribution in population genotyped in this work.